Associate AI Engineer Technical Test

May 2021

Deadline: Thursday, 27 May 2021, 1900hrs

Submissions after the deadline will not be accepted. Please submit the completed assessment early to ensure a smooth submission process.

This technical assessment consists of the following two main tasks:

1. **Exploratory Data Analysis (EDA)**
2. **End-to-end Machine Learning Pipeline (MLP)**

You are to attempt all requirements stated in the following sections and package a submission containing deliverables for the tasks specified.

In summary, the project objective is:

To analyse and predict flight delays, for both departing and arriving flights, at NY airports ("JFK","LGA"), using the provided year 2015 dataset.

1. Data Description

File download URL:
https://www.dropbox.com/s/byacfwo4e8gfo5v/Dataset%20for%20Associate%20Technical%20Test%20-%20May%202021.csv?dl=0

Data Summary:
This dataset provided information on on-time, delayed, canceled, and diverted flights, of 2015 US domestic flights operated by major air carriers.

Attributes:

Features	Description
DATE	Date of flight
AIRLINE	Airline Identifier
FLIGHT_NUMBER	Flight Identifier
TAIL_NUMBER	Aircraft Identifier
ORIGIN_AIRPORT	Starting Airport
DESTINATION_AIRPORT	Destination Airport
SCHEDULED_DEPARTURE	Planned Departure Time
DEPARTURE_TIME	WHEEL_OFF - TAXI_OUT
DEPARTURE_DELAY	Total Delay on Departure
TAXI_OUT	The time duration elapsed between departure from the origin airport gate and wheels off
WHEELS_OFF	The time point that the aircraft's wheels leave the ground
SCHEDULED_TIME	Planned time amount needed for the flight trip
ELAPSED_TIME	AIR_TIME+TAXI_IN+TAXI_OUT
AIR_TIME	The time duration between wheels_off and wheels_on time
DISTANCE	Distance between two airports
WHEELS_ON	The time point that the aircraft's wheels touch on the ground
TAXI_IN	The time duration elapsed between wheels-on and gate arrival at the destination airport

SCHEDULED_ARRIVAL	Planned arrival time
ARRIVAL_TIME	WHEELS_ON+TAXI_IN
ARRIVAL_DELAY	ARRIVAL_TIME-SCHEDULED_ARRIVAL
DIVERTED	Aircraft landed on airport that out of schedule
CANCELLED	Flight Cancelled (1 = cancelled)
CANCELLATION_REASON	Reason for Cancellation of flight: A - Airline/Carrier; B - Weather; C - National Air System; D - Security
AIR_SYSTEM_DELAY	Delay caused by air system
SECURITY_DELAY	Delay caused by security
AIRLINE_DELAY	Delay caused by air system
SECURITY_DELAY	Delay caused by security
AIRLINE_DELAY	Delay caused by the airline
LATE_AIRCRAFT_DELAY	Delay caused by aircraft
WEATHER_DELAY	Delay caused by weather
RISK_INDEX	Overall risk index for flight condition
ASSESS_CAT	Assessment category: AV - Aviation Reg, TT - Territory Threat, HE - Hazardous Environment, WH - Weather Hazard, OT - Others
SEC_STATE	Security Status
SEC_LvL	Security Level

Reference:

U.S. Department of Transportation's (DOT) Bureau of Transportation Statistics

Instructions:

The dataset can be accessed through the link below: https://www.dropbox.com/s/byacfwo4e8gfo5v/Dataset%20for%20Associate%20Technical%20Test%20-%20May%202021.csv?dl=0

Deliverable:

A **Python script** or **R script** (as part of data ingestion and preprocessing pipeline), containing filtering script to extract the dataset with the following criteria:

- The dataset should **only consist of data rows related to the stated project objective.**
- Extract all columns that are required for the necessary analysis or modeling.

You should place a copy of the final filtered **data `delay_filtered.csv`** file in a **`data`** folder in your machine learning pipeline.

Note: Include only the final filtered data in your submission, do not include the original raw data.

The total size of your submission should be limited to **20 MB maximum (refer to "Submission Format" section).**

Example:

```
├── mlp
│   ├── e.g. module1.py
│   └── e.g. module2.py
├── data
│   └── delay_filtered.csv
├── README.md
├── eda.ipynb / eda.Rmd
├── eda.pdf
├── requirements.txt / packrat.lock
└── run.sh
```

2. Exploratory Data Analysis (EDA)

Using the dataset provided, perform an EDA and create an interactive notebook that can be used present and explain the findings of your analysis. (Python or R programming language preferred.)

The report should contain appropriate and sufficient visualisations and explanations to help assessors understand how insights are derived, as well as their implications on the design of your machine learning models.

All analysis related to data preparation, input selection and feature engineering should also be included in the EDA.

(Optional), you can also provide supplementary online dashboard presentation, e.g. online dashboard using Tableau or Power BI. Please include the link to this supplementary dashboard in README.md.

Key Questions to be answered:

The EDA should include analysis to answer the following questions for NY airports ("JFK" and "LGA"): (But, not limiting to these questions.)

1. How will you define flight delay?
2. Is there any seasonal pattern for flight? (time of day, day of week, month of year)
3. Is there any seasonal pattern for flight delay? (time of day, day of week, month of year)
4. Which airline flights have more delays?
5. Which airports have more delay?
6. Is there any different flight profile between the two airports?
7. Is there any relationship between flight delay (arriving) and origin_airport?
8. Is there any relationship between flight delay (departing) and destination_airport?
9. Is there any relationship between flight duration and flight delay?
10. What are the factors that cause flight delay? Rank them.

Deliverable:

Prefered Notebook in Python or R Programming Language: an `.ipynb` or runnable `.Rmd` file with the exact naming convention as follows: - **`eda.ipynb` / `eda.Rmd`.**

Other programming languages are also allowed, but please make sure that all codes are included, with the results and findings write-up shown together with the codes. And also, all scripts included should be running properly.

Please provide a copy of the notebook (codes + results + analysis findings) in pdf format as well.

Evaluation:

You will be assessed on the clarity of visualisations, depth of insights, presentation flow and structure of your analysis.

If you are shortlisted for the interview, you are expected to be able to explain the thought processes and decisions you made throughout the analysis, and to demonstrate that you understand the underlying machine learning concepts. You may be requested to run your EDA notebook during the interview, so please make sure the EDA notebook is able to run on your laptop.

Example:

```
├── mlp
│   ├── e.g. module1.py
│   └── e.g. module2.py
├── data
│   └── delay_filtered.csv
├── README.md
├── eda.ipynb / eda.Rmd
├── eda.pdf
├── requirements.txt / packrat.lock
└── run.sh
```

3. End-to-end Machine Learning Pipeline (MLP)

Design and create a simple machine learning pipeline that will ingest/process the filtered dataset and feed it into appropriate machine learning algorithm(s), returning suitable metrics and outputs.

Deliverables:

1. A folder named `mlp` containing Python modules/classes or R scripts, or other programming language scripts. Note that Python or R programming language is preferred, but other languages are acceptable; all codes must be well structured and documented for **readability**, and **can run successfully.**

2. An executable **bash script `run.sh`** at the base folder of your submission.

3. A **`requirements.txt` file or `packrat.lock`** file, or equivalent, at the base folder.

4. A **`README.md`** file that sufficiently explains the pipeline design and its usage. You are required to explain the thought process behind your submitted pipeline in the README.

5. The README is expected to contain the following:
    a. Full name (as in NRIC) and email address.

    b. Overview of the submitted folder and the folder structure.

    c. Include information about the **programming language (with version)** used, the run environment prerequisite (including os platform and version) and the list of libraries or packages (with version) required for both the EDA and MLP in the submission.

    d. Overview of key findings from the EDA conducted and the choices made in the pipeline based on these findings, particularly any feature engineering. Also include:

        i. Summarized answers to the Key Questions listed in the EDA section

        ii. URL link for the supplementary EDA online dashboard, if any (optional)

    e. Please keep the details of EDA in the EDA notebook, this section should be a summary.

    f. Instructions for executing the pipeline and modifying any parameters.

    g. Description of logical steps/flow of the pipeline. If you find it useful, please feel free to include suitable visualization aids (eg, flow charts) within the README.

    h. Explanation of your choice of models for each Machine Learning task.

    i. Evaluation of the models developed. All metrics used in evaluation should be explained.

    j. Other considerations for deploying the models developed.

Pipeline Requirements/Specifics:

1. All codes for the MLP must be submitted. Codes submitted must be structured with well-defined functions, with good documentation.
2. **A bash script** named **`run.sh`** to run the aforementioned modules/classes/scripts.
3. **DO NOT install your dependencies in the `run.sh`**; this will be taken care of when we assess the assignment if you have created your `requirements.txt` correctly.
4. Relevant training/evaluation **metric(s) outputs to be generated** upon completion.
5. Pipeline made **easily configurable** to enable easy experimentation of different algorithms and parameters, as well as different ways of processing data (e.g. use of a config file, environment variables, or command line parameters).
6. **Python and R Programming Language are preferred for the submission.** For Python, use **only versions 3.6 and above**. For R, use **only versions 4.0 and above.**
7. Other programming languages are also allowed, but please make sure that the scripts are running properly.
8. **Please make sure that the pipeline codes can be executable successfully.** README should include clear and comprehensive setup/running instruction. (You will need to demonstrate the pipeline execution during the interview.)
9. Within the pipeline folder structure, dataset for the pipeline run must be included in the data folder, and the dataset must be filtered based on the project objective.
10. **DO NOT include the original raw data file in your submission.**

Evaluation:

You will be assessed on the quality of your code in terms of clean separation of functionality, ease of use and readability. Code reusability between the tasks will be viewed favourably.

If you are shortlisted for the interview, you are expected to be able to explain the thought processes and decisions you made throughout your code, and to demonstrate that you understand the underlying machine learning concepts.

Note for Windows users:

DO NOT submit a Windows batch (`*.bat`) script in replacement of the bash script. Use either 'Windows Subsystem for Linux (WSL)' or 'Git Bash'/'cygwin' for creation of the bash script.

Submission Format

Your work should be uploaded as a **`*.zip` file** to the designated upload link (details below). The zip file size should not exceed **20MB**. The zip file is to be named according to the following naming convention:

`<firstname>_<lastname>_<application id>_<last 5 characters of NRIC>.zip`
e.g. `john_lim_CAIE100XX_4321A.zip`

Your application id can be found in the email acknowledgement of your application.

The zip file is to have the following folder structure (as an example):

Example:

```
├── mlp
│   ├── e.g. module1.py
│   └── e.g. module2.py
├── data
│   └── delay_filtered.csv
├── README.md
├── eda.ipynb / eda.Rmd
├── eda.pdf
├── requirements.txt / packrat.lock
└── run.sh
```

Once you have packaged your submission, please access the following link to upload your zip file:
https://www.dropbox.com/sh/hp2nxmiqf9vg2g4/AADu2bgnFWBUWPOskyBIYtIRa?dl=0

If your file has been successfully uploaded, you should receive an acknowledgement email as shown below.

Note: Non-conformance to the specified conventions/formats may negatively impact your evaluation.